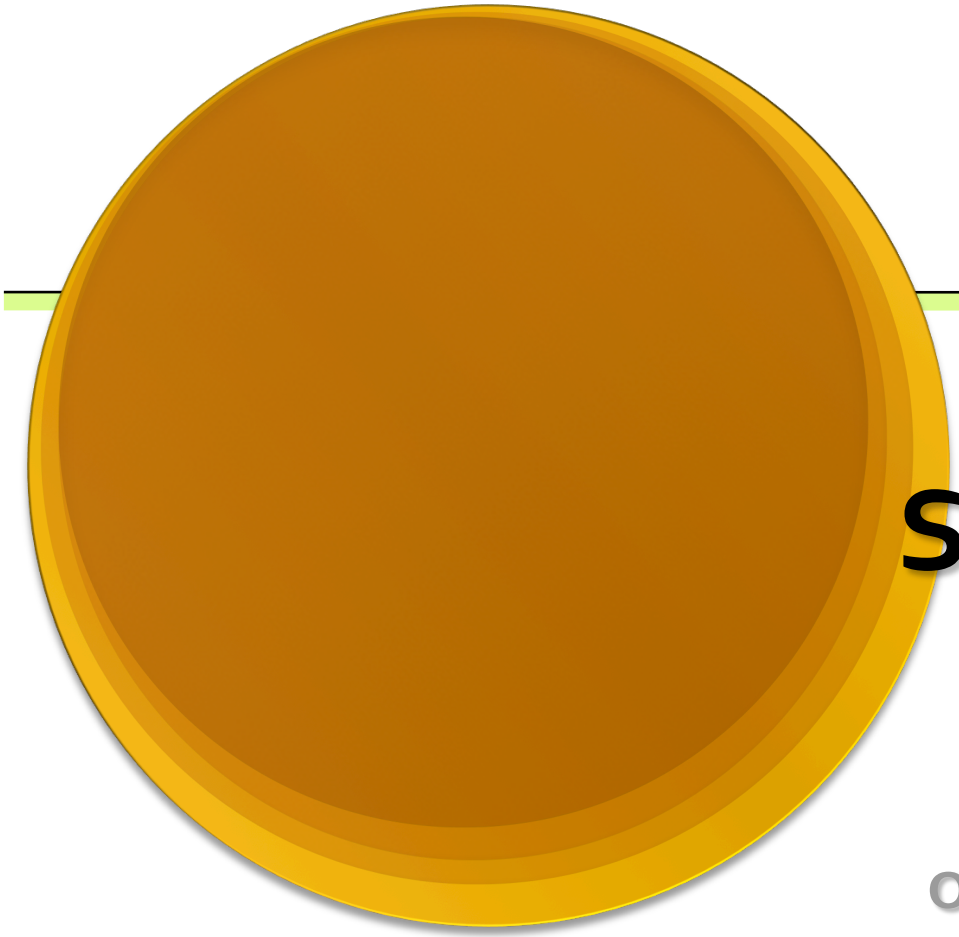


ST2334 2011/2012 Semester 2

Topic 2

Organization and Description of Data

Categorical Data Displays

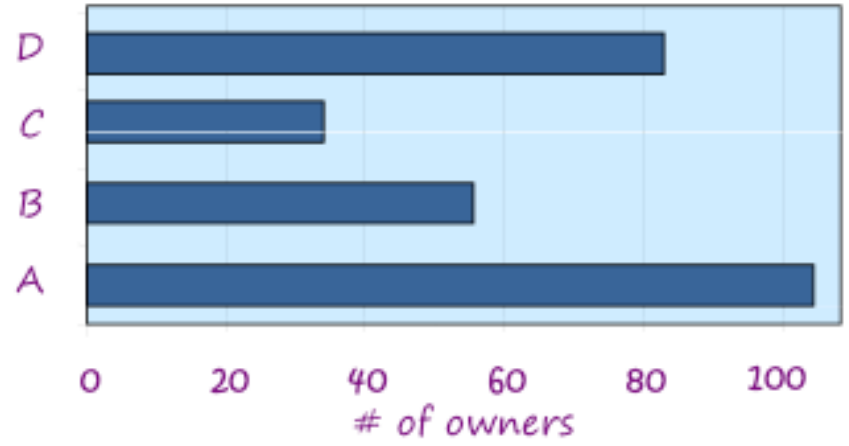
- Bar Chart
- Pie Chart
- Pareto Diagram
- The above can work for both categorical and numerical data.

Example: Laptop Brand Preference (borrowed from ST2334 2010.1)

Laptop Brand	# of owners	Percentage (%)
A	105	37.5
B	57	20.3
C	36	12.9
D	82	29.3
Total	280	100.0

Bar Chart

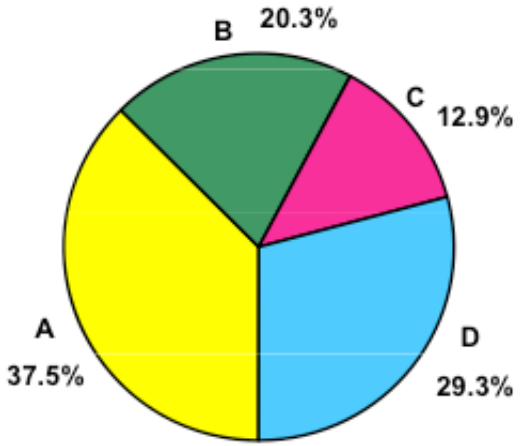
Laptop Brand	# of owners	Percentage (%)
A	105	37.5
B	57	20.3
C	36	12.9
D	82	29.3
Total	280	100.0



- Length of each bar indicates frequency or percentage
- Fixed width
- Horizontal or vertical

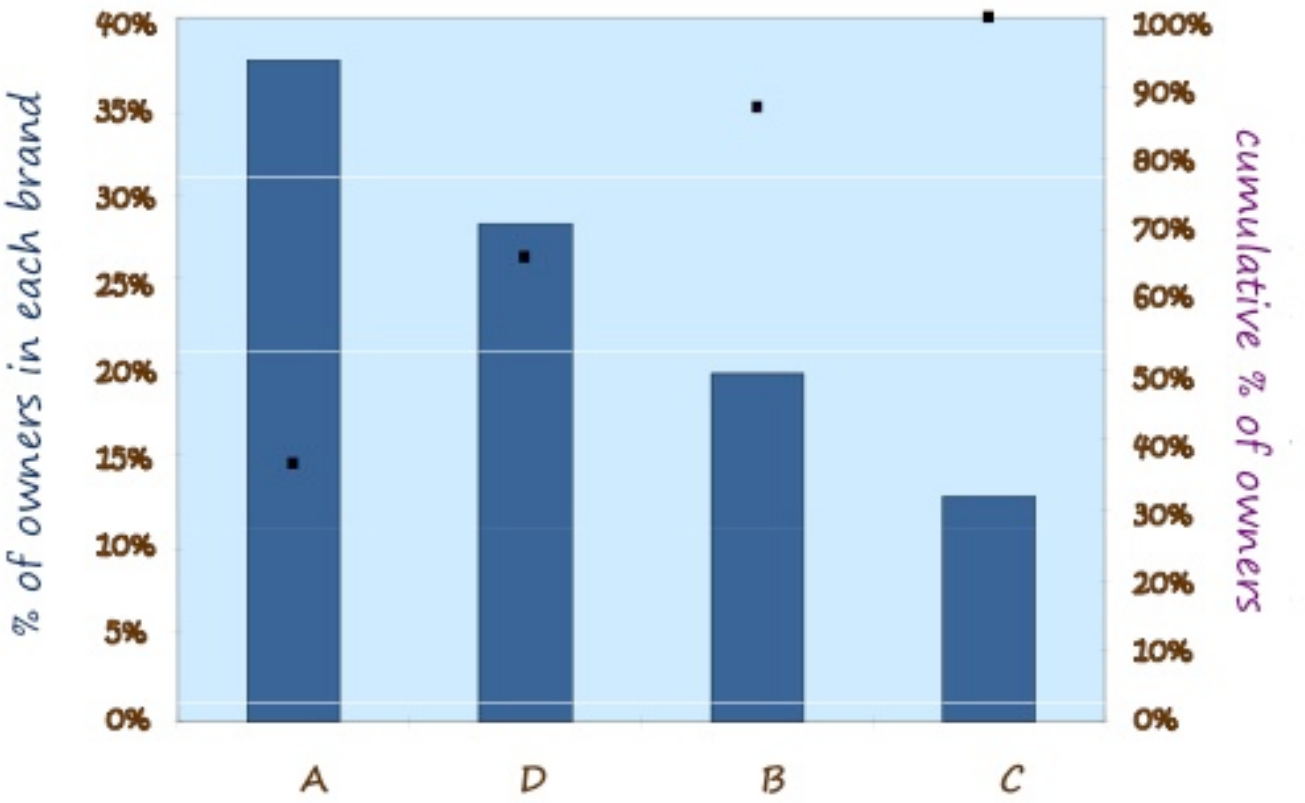
Pie Chart

Laptop Brand	# of owners	Percentage (%)
A	105	37.5
B	57	20.3
C	36	12.9
D	82	29.3
Total	280	100.0

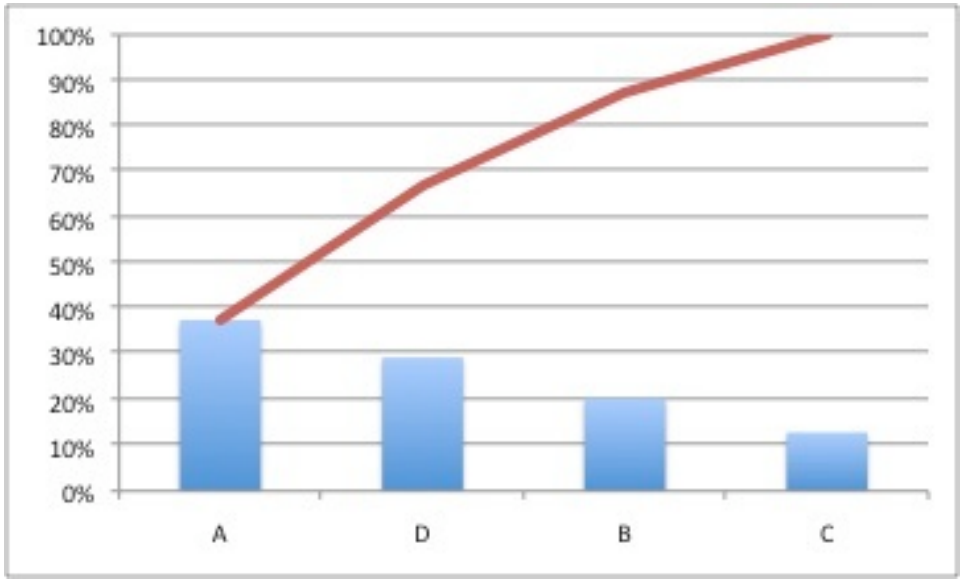


- Area (and angle) of each slice of pie corresponds to the percentage for each outcome.
- Full pie represents 100% of the observations.

Pareto Diagram (Pareto Chart)



Pareto Diagram



Pareto Diagrams

- Vertical bar chat plotted in descending order
  - "Others" category if present, plotted last
- Cumulative line graph on the same diagram
- Can use different axes for bar and line graph
  - Label carefully!
- Using common axis allow for both count and percentage to be displayed (one on each side)
- Quickly identifies major components and their cumulative share

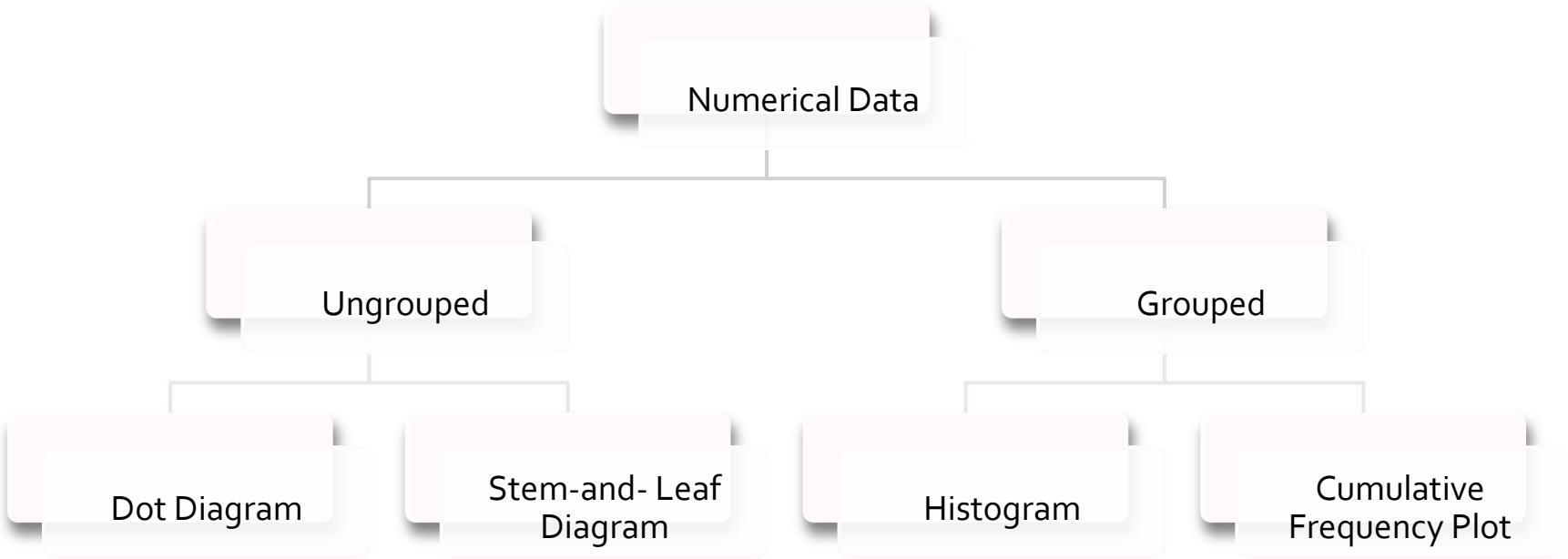
Practice!

(Units Shipments are in thousands)

Rank	Vendor	1Q11 Shipments	Market Share
1	HP	15,191	18.9%
2	Dell	10,284	12.8%
3	Acer Group	9,039	11.2%
4	Lenovo	8,172	10.1%
5	Toshiba	4,809	6.0%
	Others	33,062	41.0%
	All Vendors	80,557	100.0%

- You?

Numerical Data Displays



Dot Diagram (Dot Plot)

- A dot for each observation
- Visual summary of information when data is not too large
- May not be able to identify pattern when data is small
- Useful to find unusual features
- Example: 24, 41, 26, 24, 21, 32, 27, 38, 27, 30

Stem-and-Leaf Displays (Stemplot)

- Data sorted in ascending order
- Leading digits – Stem
- Trailing digits – Leaves
- Previous Example: 24, 41, 26, 24, 21, 32, 27, 38, 27, 30
- Sort: 21, 24, 24, 26, 27, 27, 30, 32, 38, 41
  - Can read off dot diagram too!

Two Sided Stem-and-Leaf

- Eggs produced on two farms
- Farm A
  - 50, 65, 68, 72, 74, 79, 83, 85, 87, 88, 89, 90, 93, 96, 96, 98
- Farm B
  - 53, 55, 57, 62, 65, 68, 69, 73, 76, 77, 78, 79, 82, 86, 93, 100

Grouping Data

- Can see from stem-and-leaf display that you can get meaningful data even if you ignore the details
- For exploratory data analysis (EDA), it is often helpful to group data before visualizing
- Determine
  - Number of Classes
  - Class Intervals
  - Frequency Count in each class

Grouping Data

- Number of Classes
  - Usually 5-15
- Class interval
  - Determine interval width. Equal widths?
  - Declare endpoint convention
  - No overlaps nor gaps
  - Make sure to include all observations

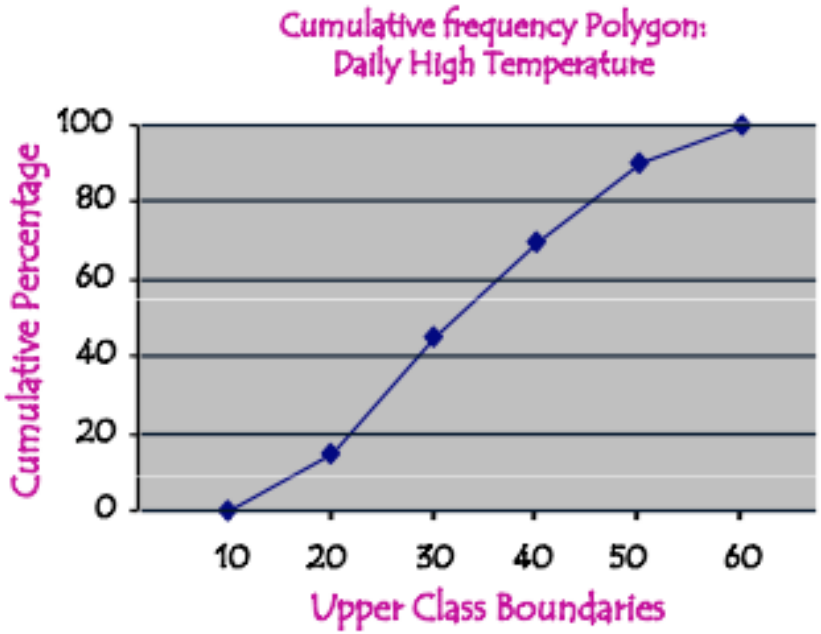
Example: Winter Temperatures

- **Data: Daily high temp (°F) 24, 35, 17, 21, 24, 37, 26, 46, 58, 30, 32, 13, 12, 38, 41, 43, 44, 27, 53, 27**

Class Intervals (Left inclusive)	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Percentage (%)
10-20	3	0.15	3	15
20-30	6	0.3	9	45
30-40	5	0.25	14	70
40-50	4	0.2	18	90
50-60	2	0.1	20	100
Total	**20**	**1**	**20**	**-**

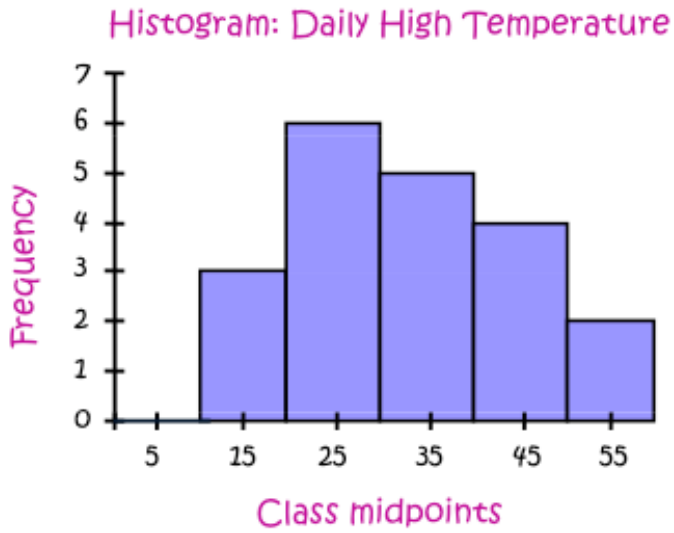
Cumulative Frequency Plot

- Horizontal axis – measurement values
- Vertical axis – cumulative frequency or cumulative relative frequency



(Frequency) Histogram

- **Similar to vertical bar chart**
  - Vertical axis is frequency or relative frequency (%)
  - Horizontal axis represents class/ group
  - Constant bar width
- **Different from vertical bar chart**
  - Bars are next to each other (scale is continuous)
  - Class interval width must be equal
  - Only for numerical data



Some notes on (Freq) Histogram

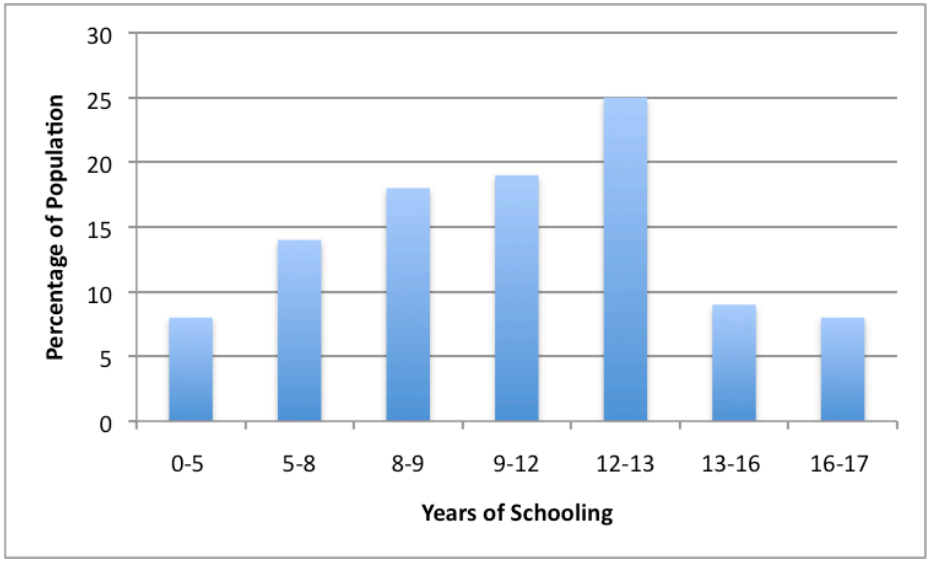
- The true intention is for the AREA of the bars to represent the frequency count.
  - Some references will have the vertical axis in frequency per horizontal axis unit. E.g. #Days per °F .
  - The two are equivalent in our case because we have the constant class interval width requirement. i.e. the shape of the histogram will not change.
  - To avoid confusion, we will NOT consider this alternative definition for this class.

Density Histogram

- Height = relative frequency / class interval width
- Area of bar represents relative frequency (percentage of data in the class interval)
- Total area = 100%
- Vertical axis unit = % per (horizontal scale unit)
- Can accommodate unequal class width
  - Useful when numerical data is naturally grouped unevenly
  - Or at times when the available data is presented that way

Example: Education Level

- United States in 1960
  - Education level for persons age 25 and over



Years of Schooling (Left inclusive)	Percentage of population
0-5	8
5-8	14
8-9	18
9-12	19
12-13	25
13-16	9
16-17	8

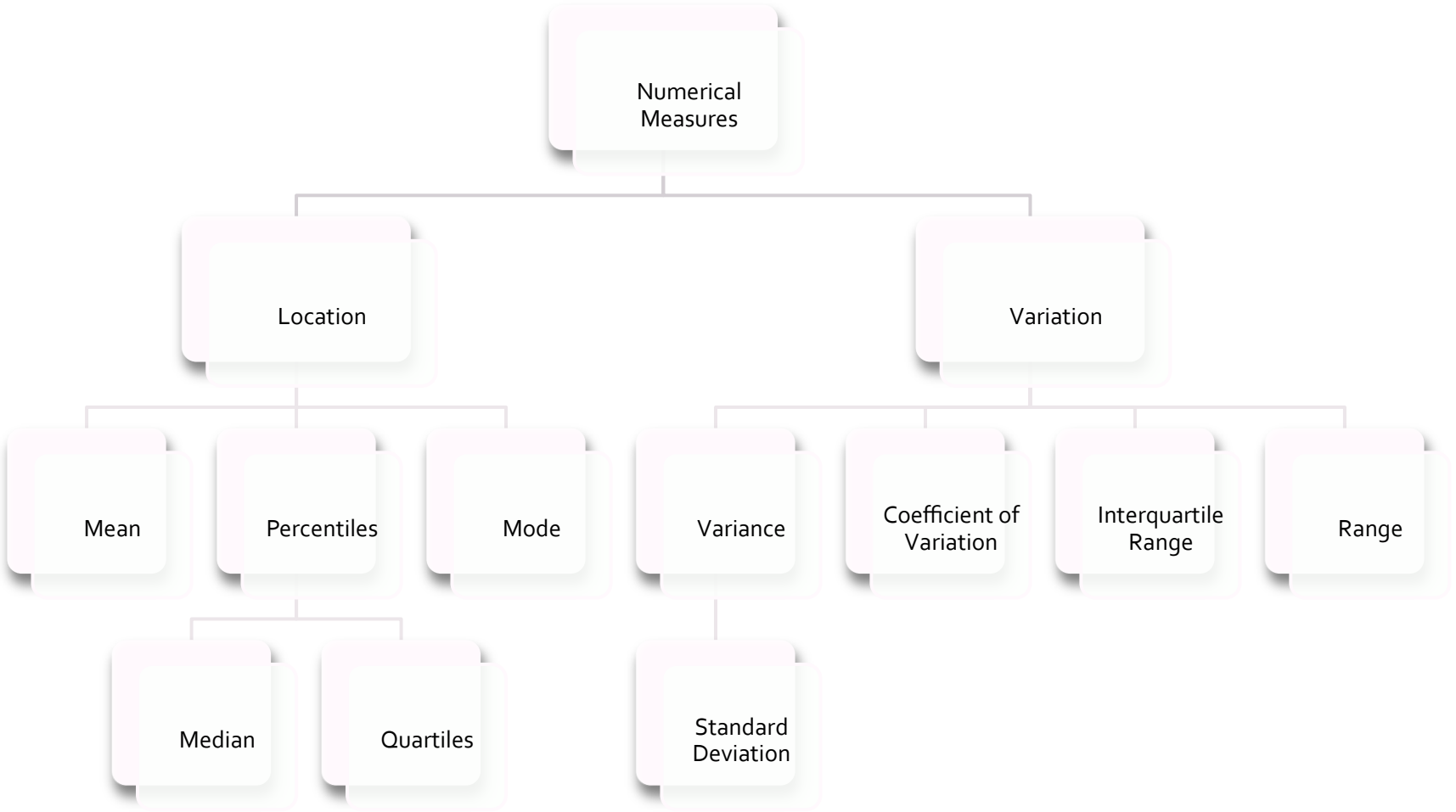
Density Histogram

- Bar chart misleading?
- Draw density histogram!
- Remember
  - Height = relative frequency / class interval width
  - Area represents percentage of population
- For this class, when we say histogram, we mean density histogram (unlike the textbook).

Steps to drawing histogram

- If classes are not defined
  - Determine number of classes
  - Range = Max – min.
  - Class interval width = range / number of classes
- Tabulate frequencies
  - Be mindful of endpoint convention
- Compute
  - Relative frequency = frequency of class / total frequency
  - Height of each bar = relative frequency / class interval width
- Draw!
  - Remember to label axes accordingly!

Descriptive Measures



Parameter vs Statistics

- Parameter – a numerical measure that describes a characteristic of the population.
- Statistic – a numerical measure that describes a characteristic of the sample.
  - Can be used to estimate parameters.

	Population	Sample
Size	N	n
Mean	μ	$\bar{x}$
Variance	σ^2	s^2

Mean

- For finite population, the population mean is
  - (we will deal with infinite population later)

$$\mu = \frac{1}{N}(x_1 + \ldots + x_N) = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

- For a sample, the sample mean is
  - Often used to estimate population mean

$$\bar{x} = \frac{1}{n}(x_1 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Sample/population mean for grouped data is

$$\bar{x}(or\ \mu) = \frac{\sum_{i=1}^{k} n_i x_i}{n(or\ N)} = \sum_{i=1}^{k} p_i x_i$$

Median

- The "middlemost" value
- For a list of *n* values,
  - If *n* is odd, median is the middle of the ordered list
    - The value at $(n+1)/2$ of ordered list
  - If *n* is even, median is any value between the two middle values of the ordered list
    - For convenience, we usually take the average of the two middle values.
    - Eg. average of values at $n/2$ and $(n/2 + 1)$ spots of ordered list.

Mode

- The value that occurs most often in finite population or sample.
  - (again, we will deal with infinite population later)

Percentiles

- 100*p*th percentile is a value such that at least 100*p*% of the observations are at or below this value, and at least (100-100*p*)% are at or above this value.
  - Simple way to find 100*p*th percentile in a list of *n* values
    - Order the *n* observations from smallest to largest.
    - Determine *np*, if integer, say *k*, take any value between *k* and (k +1) spots. If not integer, take next integer.
- Special cases
  - Median = 50^{th} percentile
  - Lower and upper quartile = 25^{th} and 75^{th} percentiles

Example: Age

- Ages of a group of friends:
  - 22, 23, 23, 23, 24, 24, 25, 25, 30, 31
- Mean?
- Median?
- Mode?
- 10^{th} percentile? Upper quartile?

On a histogram

- Mean – "balance point"
- Median – half the area on each side
- Mode – highest bar

Usually…

- As a rule of thumb, when you have a unimodal histogram with
  - a long right tail: Median < Mean
  - a long left tail: Median > Mean
- Think of simple 3 block example:

Why you so mean?

- Median describes the "center" of a skewed histogram better.
  - E.g. median income vs mean income
- Mode and median are unaffected by extreme tail values.
  - They are *robust* to outliers

Variance

- Mean squared deviation from the mean
  - Variance for finite size population

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

  - Sample Variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

    - Why (*n*-1) ?

Sample Variance

- Not really variance of the sample
  - It is just a tool to estimate the population variance
    - Mean square distance to *population* mean
    - But we don't usually have the population mean, so we plug in the sample mean instead
  - In most circumstances, by doing this plug in, we end up slightly underestimating.
    - The -1 corrects for it.
    - (We will make the conditions clear and prove this later in the course).

Grouped Data

- **Pretend each member of the group takes on a common value**

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k} n_i(x_i - \mu)^2 = \sum_{i=1}^{k} p_i(x_i - \mu)^2$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i(x_i - \mu)^2 = \frac{n}{n-1}\sum_{i=1}^{k} p_i(x_i - \bar{x})^2$$

where n_i and p_i are the number and proportion for group i, for $i = 1, \ldots, k$.

Alternative formulae for variance

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$$

- population variance with finite size

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$$

- Sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{n}{n-1}\bar{x}^2.$$

- Grouped data

$$\sigma^2 = \sum_{i=1}^{n} p_i x_i^2 - \mu^2, \qquad s^2 = \frac{n}{n-1}\sum_{i=1}^{n} p_i x_i^2 - \frac{n}{n-1}\bar{x}^2,$$

Standard Deviation

- Square root of variance
- Same units as data

$$s = \sqrt{s^2}, \qquad \sigma = \sqrt{\sigma^2}$$

Coefficient of Variation

- Relative measure of variation
  - We often care about relative precision: a weighing scale that varies about 0.1 kg works fine for measuring human weight, but is useless for measuring ingredients in a recipe!
- Standard deviation as fraction/percentage of mean (in absolute value)

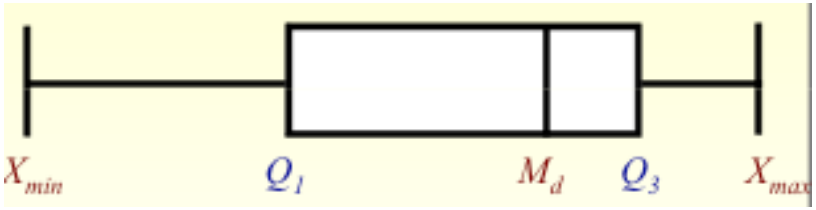
$$V = \frac{\sigma}{|\mu|} \text{ or } \frac{s}{|\bar{x}|}$$

Interquartile Range (IQR)

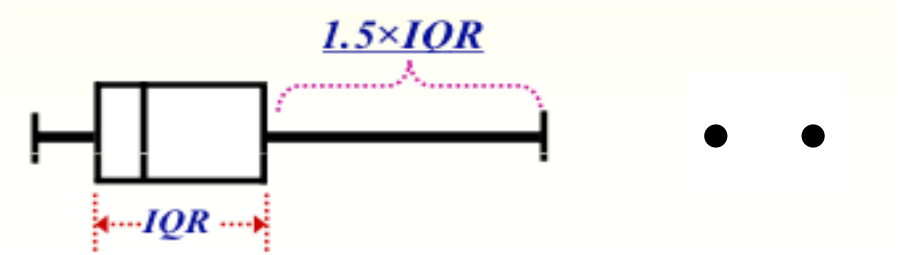
- Range = Maximum – Minimum
  - (we used this to draw histograms)
  - Also a measure of variation, but not too useful as a single large or small observation can greatly inflate its value
- Interquartile Range = Upper Quartile – Lower Quartile
  - Length of middle half of data
  - Better!

Boxplot

- Represents 5 values
  - Minimum, Lower Quartile, Median, Upper Quartile, Maximum



- Also used to display outliers, defined as further than 1.5x IQR from lower/upper quartile



Errata

- Slide 16: Added cumulative relative percentage
- Slide 29: 100*p*th 100*p*%